DOCUMENT RESUME

ED 247 257 TM 840 428

AUTHOR Ligon, Glynn; Wilkinson, David
TITLE Sizing Up Candidates for a New Achievement Test.
INSTITUTION Austin Independent School District, Tex. Office of Research and Evaluation.
REPORT NO AISD-ORE-83.57
PUB DATE Apr 84
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
AVAILABLE FROM Office of Research and Evaluation, AISD, 6100 Guadalupe, Box 79, Austin, TX 78752.
PUB TYPE Speeches/Conference Papers (150) -- Guides - Non-Classroom Use (055) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Cost Effectiveness; Elementary Secondary Education; Evaluation Criteria; Public Schools; Rating Scales; Test Interpretation; Test Reliability; *Test Selection; Test Validity
IDENTIFIERS *Austin Independent School District TX

ABSTRACT

Inspired by four recent decisions to change achievement tests used in the Austin Independent School District, the separate forms used and procedures followed have been combined into a systematic approach intended for use in future achievement test selections. A rating scale (Attachment 1) was developed to expedite a systematic comparison among possible achievement tests, and to allow a weighting of the factors to be rated according to the school system's needs. Five groups of experts (parents, teachers and principals, testing staff, central administration, and the board of trustees) have varying responsibility for rating the five factors critical to making the best choice: technical soundness; logical feasibility; instructional validity; financial affordability; and interpretational ease. The Fatal Flaw Principle (occurring when an essential factor is rated unacceptable) can eliminate a test outright, and the Shoo-In Principle (occurring when a clearly superior rating is given on a critical factor) will select a single test outright. An outline of eight procedural steps for the selection process and the contexts in which they are appropriate is attached. (BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED247257

SIZING UP CANDIDATES FOR A NEW ACHIEVEMENT TEST

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L.D. Wilkinson +

F. Holley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Glynn Ligon, Ph.D.

David Wilkinson

Office of Research and Evaluation
Austin Independent School District
Austin, Texas

Paper Presented at the 1984 Annual Meeting of the American Educational Research Association, New Orleans, Louisiana

Disclaimer

The opinions and conclusions expressed in this report are those of the author(s) and do not necessarily reflect the position or policy of the Austin Independent School District or the Office of Research and Evaluation; no official endorsement should be inferred.

TM 810 428

ORE PUBLICATION NUMBER: 83.57

Glynn Ligon, Ph.D.
David Wilkinson
Office of Research and Evaluation
Austin Independent School District
Austin, Texas

SIZING UP CANDIDATES FOR A NEW ACHIEVEMENT TEST

A major decision for a school system is the selection of a new achievement test. However, in many systems, the people charged with making that decision are doing so for the first time. Even those of us who have done it before are somewhat fuzzy about the best process to follow. What should be considered? This paper concentrates on what we as educators, administrators, and researchers should consider to get the most for our money and effort when choosing a new achievement test. The experiences from four major test changes in the Austin Public Schools, Austin, Texas, provided the base from which the forms and procedures described arose.

The Austin Public Schools have changed achievement tests at various grade levels several times recently. Each time the circumstances required a different approach to the decision. The factors which are critical to the school system somehow do not coincide exactly with the promotional information provided by the test publishers. Moreover, the factors important to the testing office do not coincide exactly with those important to other areas of the school system. The testing office, campus staff, and parents each have perspectives which make them the most appropriate ones to judge the merits of tests on different characteristics. Over the years, we have identified the factors which must be considered to represent the needs of the entire school system in the test-selection process.

In 1980, a new achievement test was introduced for grades 1 through 8; in 1982, for kindergarten; in 1982, for high school graduation competencies; and in 1983, for grades 9 through 12. As each test has been changed, we have developed instrument description forms, rating forms, and summary forms. After the new high school test was selected in 1983, the separate forms and procedures followed were combined into a unified system or approach to use in selecting future achievement tests.

A systematic approach to the selection of an achievement test can help ensure that the best decision is made--the one that returns the most useful information for the investment of money, effort, and time. To aid ourselves and others when the time

comes to select a new achievement test, this paper includes information and rating forms designed around the five factors we have found to be most critical to making the best choice. These five factors are:

1. Technical soundness,
2. Logistical feasibility,
3. Instructional validity,
4. Financial affordability, and
5. Interpretational ease.

Two principles will also be described as playing important roles in the selection process. These are:

1. The Fatal Flaw Principle and
2. The Shoo-In Principle.

The Fatal Flaw Principle comes into play when one of the five rating factors is essential for a new test. In this situation, any candidate that rates at an unacceptable level on that critical, essential factor cannot be adopted; thus, a fatal flaw exists.

The Shoo-In Principle is less precise, but a test that rates as clearly the superior candidate on a single factor that is critical and essential becomes a shoo-in in the absence of a fatal flaw.

WHO ARE THE EXPERTS?

As mentioned earlier, the perspectives and priorities of various groups of people differ when judging achievement test candidates. Figure 1 outlines the areas in which each of five groups appears to be the most appropriate to accept the responsibility for ensuring that the correct choice is made. This is not to say that all groups are not to consider all factors, but the group(s) with the major responsibility for each factor is(are) indicated.

WHAT ARE THE STEPS FOR SELECTING AN ACHIEVEMENT TEST?

Figure 2 outlines the steps which are usually taken when an achievement test is selected. There appear to be three contexts within which test are chosen. These three contexts are:

1. Any test may be chosen,

2. One test is clearly the best, and

3. One test is mandated.

All three of these situations have been encountered in Austin, and we have found that following only the steps necessary saves everyone time from pursuing nonoptions.

WHAT IS THE RATING SCALE?

Attachment 1 presents the rating scale, titled "Factors for Rating Achievement Tests." The rating scale has two intended purposes:

1. To expedite a systematic comparison among achievement tests, and

2. To allow a weighting of the factors to be rated according to the system's needs.

The rating scale contains three components:

1. Subfactor ratings,

2. Factor ratings, and

3. Overall weighting for the test.

Subfactors are the subsidiary considerations which, taken as a whole, make up each of the five factors. Subfactors are assigned ratings by the individual or group using the rating scale on the following basis:

5 = Adequate
4 = Mostly Adequate
3 = Partly Adequate, Partly Inadequate
2 = Mostly Inadequate
1 = Inadequate

For example, if under the factor of technical soundness, the reliability of the candidate test is judged by the rater to be "mostly adequate," the rater would assign a rating of 4 to this subfactor. The number of subfactors under each of the factors differs. However, since an arithmetic average of the subfactor ratings is taken to arrive at the five factor ratings, each subfactor component is equally represented in the factor ratings. In cases where the average is a decimal fraction, the average should be rounded to the nearest whole number. As will

be noted from Attachment 1, several types of reliability should be considered by the rater. In such cases where the subfactor is not a unitary concept, the rater may choose to assign ratings to each of the subcomponents under the subfactor and first average these, or may simply consider these subcomponents in making an overall subfactor rating.

It should be noted that these subfactor ratings are a matter of subjective judgment on the part of the rater, based on the rater's experience and familiarity through study of the test under consideration. In fact, subfactor ratings will probably differ from individual to individual and from group (administrators, teachers, parents, etc.) to group. Therefore, it is to the benefit of the school district to obtain multiple ratings of a test for comparison and as a basis for further study and discussion should the differences be extreme.

However, it is in this connection that the Fatal Flaw Principle may usefully be employed. If any of the five factor ratings, made up of the averages of the subfactor ratings, is less than 3, i.e., mostly inadequate or inadequate, the test should be dropped from consideration. This is a useful procedure which helps to narrow the range of candidate tests and which serves as an anchor for the different rating groups.

The third component of the rating scale, an overall weighting of each of the five factors, is an additional mechanism for separating inadequate tests from adequate tests based on judgments of the importance of each of the five factors. The overall weightings should be treated as percentages summing to 100. For example, the logistical feasibility of a test may be of paramount importance to a district. In this event, the rater might assign a 60% weight to this factor and, perhaps, a 10% weight to each of the remaining four factors. To arrive at the weighted factor ratings, then, the rater would multiply the factor ratings by the percentage weights. In the example, if the factor rating for logistical feasibility was 3, the weighted factor rating would equal 180. Note, however, that after such weighted factors have been obtained, the only valid comparisons are between tests, not between factors on the same test. Should a rater desire to compare factors on the same test, the unweighted factor ratings should be used.

WHAT ARE THE FIVE FACTORS?

Technical Soundness

This first factor is the one which, initially, is of the most importance to the district testing office, and of lesser concern to the other consumers. Ultimately, however, it is the base upon which the whole testing enterprise will rest because, sooner or later, given the high visibility which test scores have

assumed for an accountability-conscious public, the test which is chosen must stand up to the critical review of the other consumers. Therefore, this factor must be given thorough consideration, even to the extent of disqualifying from the outset some tests which are apparently attractive from the standpoint of the other four factors. In fact, as embodied in the fatal flaw principle, each of the five factors contains features which may disqualify a test from consideration despite the test's strengths in other areas. This paper does not purport to be an exhaustive examination of the technical bases for evaluating the psychometric properties of a test since these aspects are covered at length elsewhere. However, the rating scale should be a workable checklist touching on the major criteria.

Logistical Feasibility

This factor attempts to address those features of a test concerned with the logistics of the actual administration of the test. One consideration is the levels of the test which are available. Is there a level available for each grade to be tested, or are several grades to administered the same test level? In either event, there should be good articulation across test levels.

A second consideration is whether there are alternate forms of the test available. Alternate forms are desirable for several reasons:

1. If the same form of the test is given repeatedly, particularly if the same level is given to different grades over the course of several years, as was the case with AISD's high school achievement test, the test takers become familiar with the test merely from its repetition. In this event, it is useful to have at least one alternate form of the test to ensure that reliable test results are obtained. In Austin, two forms of the high school achievement test were alternated annually at each high school. In this way, the test takers were presented with a slightly different test in alternating years.

2. Familiarity with the test also tends to promote a shift in instructional practices toward the content of the test. An alternate form of the test helps ensure that the instructional focus does not become too narrow, but must instead remain sensitive to the slight differences from one form of the test to the other.

3. In the instance where cheating on the test is suspected, an alternate form of the test may

be administered for comparison purposes.

Functional-level testing may also be a desirable characteristic of a test.

Another consideration is the amount of time the test requires for administration. With a premium on instructional time, it is desirable for the test to take as little time as possible away from instruction. Hence, a test whose subtests are shorter translates to fewer days over which the test must be administered.

Finally, ease of scoring is an important consideration in terms of a test's logistical feasibility. In smaller school districts, where scoring is done manually, a test which is readily hand scored and which quickly generates derived scores is preferable to a test with complex scoring procedures which may necessitate sending the answer documents to the test publisher for scoring, at extra cost to the district and with a consequent time delay in receiving the scores. In larger school districts, which have the capability to score the test "in house," a test with the requisite conversion tables and other technical documentation readily available is essential.

Instructional Validity

This is a factor of prime concern, since unless the test covers the instructional areas in which information is required, it cannot be valuable, even if it has excellent technical soundness. In this regard, the match between the test and the curriculum must be as close as possible. A test that matches the system's curriculum very well might be a shoo-in candidate. It should be noted that a perfect match between a district's curriculum and the content of a commercially published test is unlikely. Therefore, an alternative for a district is a locally developed test which can be constructed which would ensure a closer match. However, this is an expensive and time-consuming alternative. An additional drawback is the lack of national norms with which to compare district achievement.

Related to the match with curriculum is the consideration of the terminology used by the candidate test. If the language used in the administration directions and in the test items is a noticeable departure from that customarily used by teachers in their everyday instruction, some provisions have to be made early in the school year, well before the test administration, to acquaint the students with the test's terminology.

Finally, an important consideration, perhaps of overriding importance, is the utility of the test results for instruction. Apart from the accountability function of the test for curriculum decisions, the test must generate useful information for teachers to use in the instruction of individual students. To this end, the test must contain information for districts to create skill profiles of individual student strengths and weaknesses.

Financial Affordability

This factor is an important and sometimes overlooked concern. Even if a test is excellent in terms of its technical soundness, logistical feasibility, and instructional validity, it must be affordable by the district. Financial conside ations include:

1. Start-up costs. The district must make a large initial outlay of funds to purchase teachers guides, scoring keys, and the like for annual use. It must also purchase reusable booklets, if the test is for grade levels above the early primary, for students who mark on a separate answer sheet.

2. Annual costs. For students in the early primary grades who cannot use a separate answer sheet and must mark directly onto the test booklet, the district must purchase replacement booklets annually. It must also purchase additional teacher guides and other testing materials as these become delapidated due to ordinary use or due to the rigors of shipping.

Interpretational Ease

This factor is related to the utility of the test for generating interpretable results. A variety of scores should be available, including an overall composite score, scores for each of the major domains tested, e.g., a total mathematics score, and scores which can be grouped according to each of the subskills tested, e.g., number sense. The test should have norming dates useful for annual achievement comparisons and for major projects such as Chapter 1. The test battery should include tests in areas of concern to school districts, including newer areas such as life skills and computer literacy. Test scores should allow comparisons both on a longitudinal and a cross-sectional basis.

CONCLUSION

Sometimes we think that a school system should never change achievement tests--especially if we know the implications of replacing one test with another. However, once past the hurdle of committing to a change, learning from the experience of others can make the selection process much more efficient and productive. In this paper, we have outlined and described the following aspects of the process of selecting an achievement test:

1. The factors to consider,

2. The people with the greatest responsibility to ensure adequacy for the new test on each factor,

3. The Fatal Flaw Principle, which eliminates a test outright,

4. The Shoo-In Principle, which selects a single test outright, and

5. The context within which a test is selected.

Is it pretentious of us to have laid out such an elaborate process for others to follow without having first put it to the test, so to speak? Of course it is. Our purpose, we must admit, was to document our ideas for ourselves while the emotions of our recent test changes are still fresh. We do intend to follow the suggestions of this paper the next time we change one of our achievement tests. Until then, if any reader takes our plan to heart, let us know how it works.

	Technical Soundness	Logistical Feasibility	Instructional Validity	Financial Affordability	Interpretational Ease
Parents					X
Teachers and Principals		X	X		X
Testing Staff	X	X	(X)	(X)	(X)
Central Administration		(X)	(X)	X	(X)
Board of Trustees				X	X

Figure 1: WHO ARE THE EXPERTS? Who has major responsibility to ensure adequacy on each factor for the new achievement test?

Figure 2: STEPS FOR SELECTING AN ACHIEVEMENT TEST

STEPS	CONTEXT		
	Any Test May Be Selected.	One Test is Clearly the Best.	One Test is Mandated.
Compile data on all available tests (include locally made test option).	X	X	
Appoint a review committee.	X	(X)	
Committee rates all available tests.	X	(X)	
Committee reviews test items.	X	(X)	
Committee makes a recommendation or provides comments.	X	(X)	
Appropriate administrative office makes recommendation.	X	X	
Key groups review recommendation or mandate.	X	X	X
Final recommendation goes to the Board of Trustees.	X	X	X

FACTORS FOR RATING ACHIEVEMENT TESTS

Test: _____

Directions: Assign a numerical rating from 1 to 5 to each of the factors which follow. In this scale:

5 = Adequate
4 = Mostly Adequate
3 = Partly Adequate, Partly Inadequate
2 = Mostly Inadequate
1 = Inadequate

Factor ratings will be based on an average of subfactor ratings.

Ratings (1-5)	Factors
_____	I. Technical Soundness
	(Primary raters: Testing Staff)
____	A. Reliability
	1. Test-Retest
	2. Internal Consistency
	3. Correlation with Other Tests/Forms
____	B. Validity
	1. Divergent
	2. Factorial
	3. Concurrent
	4. Predictive
____	C. Norms
	1. Empirical Norms
	2. Critical Norming Dates
	3. Norm Sample
	a. National Representation
	b. Size of Sample
	c. Subgroup Norms
	1. Urban
	2. Regional
	d. Consistencies
	1. 50th %ile = GE for time of testing
	2. +1.0 GE/year growth at 50th %ile
	3. >1.0 GE/year growth above 50th %ile
	4. <1.0 GE/year growth below 50th %ile
	5. Standard score growth rate logical
____	D. Fairness
	1. No Sex Bias
	2. No Ethnic Bias

II. Logistical Feasibility

(Primary raters: Teachers and Principals, Testing Staff)
(Secondary raters: Central Administration)

____ A. Levels Available per Grade Span

____ B. Alternate Forms Available

____ C. Out-of-Level Testing
1. Booklet Adaptations
2. Administration Differences Across Levels

____ D. Testing Schedule
1. Days
2. Time per Test

____ E. Ease of Scoring
1. Manual Scoring
2. Publisher's Scoring Service
3. District's Machine Scoring

____ F. Format
1. Physical Layout of Items
2. Print
3. Graphics and Illustration

____ G. Clarity of Directions
1. For Test Administrators
2. For Students

____ H. Availability of Student Practice Materials

____ I. Training Requirements for Administrators

____ J. Other Language Editions

____ K. Editions for the Handicapped

_____ III. Instructional Validity

(Primary raters: Teachers and Principals)
(Secondary raters: Testing Staff, Central Administration)

___ A. Areas Tested

___ B. Match with Curriculum
1. Content
2. Skills Levels
3. Terminology

___ C. Utility of Test

_____ IV. Financial Affordability

(Primary raters: Central Administration, Board of Trustees)
(Secondary raters: Testing Staff)

___ A. Start-Up Costs
1. Reusable Booklets
2. Teacher Guides

___ B. Annual Costs
1. Disposable Booklets
2. Answer Sheets
3. Scoring
4. Reporting

___ C. Adjusted Annual Per Pupil Cost: [(Start-Up Costs) = (Annual Costs x Years Life Expectancy of Test] - Years Life Expectancy of Test

_____ V. Interpretational Ease

(Primary raters: Parents, Teachers and Principals, Testing Staff, Central Administration, Board of Trustees)

____ A. Scores Available
1. Overall Composite
2. Major Domains
3. Skills Areas

____ B. Norming Date

____ C. Test Areas

____ D. Comparability
1. Past Scores
2. Other Districts/Groups

Test Rating Summary

Factor	Weight	x	Rating	=	Total
I					
II					
III					
IV					
V					
	100				Overall Rating

Weight: Divide 100 up among the five factors (for example: 10, 30, 20, 10, 30) to represent the relative importance of each in the decision.

Suggested Reading

Hoepfner, R., Strickland, G., Stangel, G., Jansen, P., & Patalino, M. (1970). CSE Elementary School Test Evaluations. Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation.

Hoepfner, R., Stern, C., & Nummedal, S. G. (1971). CSE-ECRC Preschool/Kindergarten Test Evaluations. Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation and the Early Childhood Research Center.

Hoepfner, R., Conniff, W. A., Hufano, L., Bastone, M., Ogilvie, V. N., & Hunter, R. (1974). CSE Secondary School Test Evaluations: Grades 11 and 12. Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation.

Hoepfner, R., Conniff, W. A., McGuire, T. C., Klibanoff, L. S., Stangel, G. F., & Lee, H. B. (1974). CSE Secondary School Test Evaluations: Grades 9 and 10. Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation.

Hoepfner, R., Conniff, W. A., Petrosko, J. M., Watkins, J., Erlick, O., Todaro, R. S., & Hovt, M. F. (1974). CSE Secondary School Test Evaluations: Grades 7 and 8. Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation.

Ligon, G. (1982, March). Warning! Iceberg! A checklist of issues related to changing achievement tests. Paper presented at the annual meeting of the American Educational Research Association, New York, N.Y.

Madaus, G. F., Airasian, P. W., Hambleton, R. K., Consalvo, R. W., & Orlandi, L. R. (1982). Development and application of criteria for screening commercial, standardized tests. Educational Evaluation and Policy Analysis, 4 (3), 401-415.

Matter, M. K., & Ligon, G. (1983, April). Preparing students for standardized testing: Everybody's business. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

A copy of this report may be obtained

from the address below. Pub. No. 8357

OFFICE OF RESEARCH AND EVALUATION, AISD,
6100 GUADALUPE, BOX 79, AUSTIN, TX. 78752